DOCUMENT RESUME

ED 152 793 TM 006 887

TITLE A Report on the Feasibility and Design of Alternative Approaches to Evaluating Preceptorships Exclusive of Longitudinal Tracking. Final Report-Part II.

INSTITUTION American Medical Students Association Foundation, Rolling Meadows, Ill.

SPONS AGENCY Health Resources Administration (DHEW/PHS), Bethesda, Md. Bureau of Health Manpower.

PUB DATE 30 Sep 76

CONTRACT BHM-231-75-0606

NOTE 68p.; For related documents, see TM 006 886 and 888

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.

DESCRIPTORS *Career Choice; Clinical Experience; Community Size; Control Groups; Cross Sectional Studies; Data Collection; *Evaluation Methods; *Feasibility Studies; Field Experience Programs; Geographic Regions; Higher Education; *Medical Students; Predictor Variables; Primary Health Care; Program Evaluation; *Research Design; Research Methodology; *Research Problems; Specialization; Validity

IDENTIFIERS *Preceptorship, Ex Post Facto Design

ABSTRACT

The original version of this report dealt with the feasibility of a longitudinal design for studying the effects of preceptorships on medical students' career choices. The present document supplements the original report, focuses on various non-longitudinal designs, and assesses their suitability for providing the type of information required for evaluating the impact of preceptorship experiences on certain aspects of medical career paths. Each design is evaluated in terms of its ability to provide reliable information about three outcomes: speciality, geographic area, and community size choices. This report includes seven sections: (1) the various problems of control in evaluation designs, including the categories of variables which must be taken into account when preceptorship impacts on career choices are being studied; (2) retrospective, or ex post facto designs; (3) cross-sectional designs; (4) the possibilities of pooling existing data in a centralized computer file; (5) the possibilities of constructing an evaluation package for use by several preceptorship programs; (6) some general issues in choosing an evaluation design; and (7) additional questions regarding the impacts of preceptorships which might be worth studying. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152793

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: TM HE
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view. CG

A Report on the Feasibility and Design of Alternative Approaches to Evaluating Preceptorships Exclusive of Longitudinal Tracking

Submitted by the

American Medical Student Association (AMSA) Foundation

FINAL REPORT - Part II

TM006 887

September 30, 1976

FOREWORD

This report contains the second portion of the work performed by the contractor on Contract #231-75-0606. In April the contractor submitted to the Bureau of Health Manpower its report on a longitudinal approach to evaluating preceptorships. This report reflects the contractor's work requested by the Bureau on other alternative approaches to evaluating preceptorships. The two reports together comprise the contractor's total effort for complying with that portion of the Scope of Work of Contract #231-75-0606 addressing the Longitudinal Feasibility Study.

TABLE OF CONTENTS

Introduction 1

I. The Importance of Control in Evaluation Studies. 2
The Concept of Control 3
Categories of Variables Which Must be Controlled in Studying Preceptorships 3
The Need for Control 4
Problems Arising from Inadequate Control 6

II. Retrospective Designs. 9
General Weaknesses 9
Retrospective Data Currently Available 12
Potential Sources of Additional Retrospective Data 13

III. Cross-Sectional Designs. 15
General Weaknesses 15
Cross-Sectional Data Currently Available 17
Evaluation of Existing Evaluations 20
Cross-Sectional Designs Involving Additional Data Collection: Some Considerations. 26

IV. Compiling Data From Existing Studies 30
The Problem of Comparability 31
Identifying Program Differences. 33

V. Use of Packaged Designs. 35

VI. Some General Issues in Selecting an Evaluation Design. 40
Reasons for Disagreements Involving Weak Designs 41
Objections to Sound Designs. 43
Improper Use of Weak Designs 45

VII. Other Types of Impacts Which Could be Considered 46

References 49

Table 1 - List of Schools and Programs From Whom Information on Preceptorships Was Received 51

Table 2 - Tabulation by Various Categories of Schools Listed in Table 1. 55

Table 3 - Descriptions of Data Obtained by Those Evaluation Studies Which Collect Career Preference Information. 56

Table 4 - Evaluation Designs of Programs Collecting Career Preference Data. 58

Table 5 - Selected Deficiencies in Evaluation Design Rank Ordered by Degree of Seriousness 59

Table 6 - Type of Data Collected by Four Different Programs on Three Outcome Variables. 60

Table 7 - Major Design Options for Evaluating Preceptorship Impacts. . 61

INTRODUCTION

The original version of this report, delivered to the Bureau of Health Manpower under Contract 231-75-0606, dealt with the feasibility of a longitudinal design for studying the effects of preceptorships on medical students' career choices. The present document is a supplement to that original report and focuses on various non-longitudinal designs and attempts to assess their suitability for providing the type of information required for evaluating the impact of preceptorship experiences on certain aspects of medical career paths.

Through discussions with Bureau personnel, parts of this effort have been quite clearly specified, while others have been left open-ended. The outcomes to be addressed by the various designs have been specifically designated as: specialty, geographic area, and community size choices. No constraints have been placed on the designs to be considered, either in terms of complexity, size of sample, political considerations, or costs.

The choice of a methodologically appropriate evaluation design depends on (a) the questions to be answered, and (b) the specific situations, circumstances, and programmatic contexts in which those questions are being asked. For purposes of this report, it has been necessary to place some heuristic constraints on the items in "b". The intent has been to achieve a compromise between a too narrow discussion and a too open-ended one. To this end, issues are discussed which it is hoped can aid the Bureau in choosing among various categories of evaluation designs. Specific designs are considered in this report only when they serve to illustrate existing evaluation efforts or a general principle of design.

Prior to the writing of this report, a review was made of the evaluation efforts currently under way at various schools. The review was conducted for the proposes of (a) determining whether there existed any data which could be used for evaluating the preceptorship program outcomes listed above, and (b) determining whether any of the extant evaluation efforts could be packaged for use in a evaluation study of national scope. The program information available for this review consisted of selected correspondence from the files of (a) the National Health Council, and (b) the authors of a preceptorship handbook soon to be published by the AMSA Foundation.

Table 1 lists schools and programs from whom information on preceptorships had been requested by the sources just mentioned. It also contains selected information about these programs. Table 2 shows tabulations of some of the basic kinds of information which were available from the programs listed in Table 1 and which comprise the basic information pool on which much of this report is based.

The present report is organized into seven sections. Part I considers various problems of control in evaluation designs, including the categories of variables which must be taken into account when preceptorship impacts on career choices are being studied. Part II considers retrospective designs; Part III covers cross-sectional designs; Part IV examines the possibilities of pooling existing data in a centralized computer file; Part V considers the possibilities of constructing an evaluation package for use by several preceptorship programs; Part VI focuses on some general issues in choosing an evaluation design; and Part VII concludes with a brief account of some additional questions regarding the impacts of preceptorships which might be worth studying.

I. THE IMPORTANCE OF CONTROL IN EVALUATION STUDIES

There are several major topics to be discussed in this section: identification of factors for which evidence indicates their influence on career preferences; the need for and nature of controls for these factors; and the problem of determining what effect a lack of adequate control may have on study results and interpretation. This section is not intended to be a treatise on experimental and evaluation design principles, but it is important that policy-makers who are concerned with making decisions about which programs to implement or to continue to fund be aware of the nature of the issues involved here and their implications for policy and funding decisions. The behavioral sciences are "soft" in the sense that they can rarely provide definite, black and white answers. Even among behavioral scientists there is disagreement about methodologies, especially in the area of some of the weaker evaluation designs. In other areas, there is surprising agreement, however, and the disagreements do not mean that these sciences have

little to offer the policy-maker. What they do mean is that policy-makers can benefit most from the inputs of behavioral scientists if they have at least a rudimentary understanding of the shortcomings of social science methodology and of the possible consequences of these shortcomings.

The Concept of Control

The concept of control in the context of evaluation design is a crucial one which is central to assessing the value of the informational return provided by any evaulation effort. "Value" as used here refers to the degree to which the empirical findings of an evaluation study correspond to the reality being studied. The "reality" defined for purposes of study is in itself a matter of critical concern for those making use of the study results, but is not an appropriate consideration here (see Weiss, 1975, pp. 19-23 for a further discussion). In medical education there are many experiences and factors which act upon students and have the potential for influencing their future behavior, and specifically career choices. To determine the impact of any one of these influences requires that we be able to control the remaining factors which are thought to influence students' career preferences. Control as used here refers to the ability to hold constant these other factors or to measure their variation in such a way that their influences can be statistically separated from the influence of (in this case) preceptorship experience.

Categories of Variables Which Must be Controlled in Studying Preceptorships

The factors which are known to influence medical students' career preferences and choices are many, as might be expected from the complexity and intensity of medical education. Repeated investigations over the past decade and even earlier indicate that among the categories of variables found to influence student career dispositions are: student background; ability factors; personality factors; and experiential factors (Otis et.al., 1974). This list, of course, refers only to categories of variables found to influence career paths. For purposes of investigating the influence of preceptorship experiences on students' career interests it is necessary to identify specific variables within each category, and this is where the complexity of the control problem becomes evident. Among the background factors found

to be relevant in predicting career preferences, for example, are socio-economic status of parents, size of community of origin, number of siblings, sex, and type of college attended (Otis et al., 1974). No attempt will be made here to list all of the variables in each category which have been found relevant to career path choices. The main point to be made is that each of the variable categories mentioned above contains numerous variables, each of which has been empirically found to be related to career preferences. Thus, it is not sufficient to control just any variable which could be classified as a background or an experiential variable. It is those variables which are apt to influence career decisions which must be controlled.

The Need for Control

Control is an essential component in any evaluation if one is attempting to assess program effects or outcomes. As Houston states, the evaluation design must

> ...permit the comparison of units exposed to the program with a control group which is not. The evaluator wishes to interpret measured differences between these groups as program impact; but a necessary condition for such inference is the absence of prior differences between the two groups....randomization is an elegant means of equating groups (Houston, 1972, p. 53).

Randomization and other designs aimed at similar control purposes have not found favor in evaluation studies for a variety of reasons. Among the objections which have been raised to the use of sound designs are: their greater cost; the greater time required for results to come in; ethical questions arising from the manipulation of people in assigning them to treatment and control groups; political, organizational, and programmatic constraints which make it difficult to employ adequate control groups; and concerns from interested parties about what the study results will show. These objections will be considered at appropriate points throughout this report as our discussion warrants.

Those who are in positions of needing to use the results of evaluation studies frequently take the position that the "best" designs (which are often mistakenly presumed to be the most costly; cf. Lumsdaine and Bennett, 1975, p. 551) are unnecessary or not feasible for evaluation of social programs.

Unfortunately, such a viewpoint is rather myopic, for as Gilbert et al. state,

> The purpose of randomized controlled field trials, like that of the other kinds of evaluative investigations we discuss, is marshalling information to aid decisions. People tend to think of this device as especially important for the scientist rather than for the policy maker. This leaves us with the idea that a scientist needs a very fine tool to get information about a process, but the policy maker should be content with uncertain information. Although sometimes this may be true, one could argue that exactly the reverse holds: the policy maker needs not only fine tools, but also good efforts to develop better tools. This seems especially true in policy problems where large numbers of individuals are involved and large amounts of money are at stake or when lives and careers are being substantially affected. (Gilbert et al., 1975, p. 52)

The better evaluation designs, i.e., those providing more certain information, are of such a nature that they must be initiated at the beginning of a program's implementation, and planning for the evaluation study should be an integral part of the preliminary program planning. This is one of the reasons behind Hilton and Lumsdaine's statement that

> Planning and initiation of evaluation activities, including design of studies and procedures for collection of data, should be initiated at the very outset of program planning, rather than being deferred until after the program is already under way. (Hilton and Lumsdaine, 1975, p. 394)

Only when evaluation is planned from the beginning is it possible to utilize sound designs to assess program impacts. After programs have been under way for some time it is difficult to obtain the necessary control mechanisms which are essential for determining program impacts with any degree of certainty, at least on a national scale.

It is not possible to determine from this time point the reasons why the initial funding of preceptorships was not accompanied by an evaluation effort. Nevertheless, the consequences of this omission are clear: the designs which can be implemented now to yield results in a short time are much weaker than those which could have been employed earlier in the preceptorship funding experience of the Bureau.

This last point leads us to another important consequence which will be discussed here, and this is that the evaluation designs to be considered in this report are among the weaker types of designs. It should be recognized, therefore, that when the designs presented are ranked later by the adequacy

(certainty) of the information returned the "best" of them may not be worthy of the name "design" (Houston refers to them as "pre-experimental designs" or worthy of consideration by policy makers).

Problems Arising from Inadequate Control

The problem with the weaker designs is one which must be confronted directly by the policy maker wishing to make use of the findings from evaluation studies, and this problem is that the degree of confidence in or soundness of the findings is reduced considerably. "Soundness" as used here refers to the adequacy of the justification for accepting the findings of the evaluation, and not to the political, bureaucratic, or popular support which may become aligned for or against the results of the study. The basic problem with weak designs is that the uncertantities are great enough that one cannot know whether the results observed were due to the program in question, whether they would have occurred without the program or even whether the results might not be an artifact of the weak design itself. Some of these problems ensuing from design weaknesses will be discussed later in conjunction with specific examples, as will some of the broader consequences of weak evaluations. Here we will merely note some general types of problems which are likely to be encountered across several categories of designs.

One of these problems is self-selection. Whenever individuals enter a program of their own volition, extreme difficulties in interpretation of results are encountered no matter how many "comparison" groups are present or how many variables are being controlled statistically. In the case of preceptorships, for example, those individuals voluntarily entering preceptorship programs already have some interest in primary care or shortage settings,and it is virtually impossible to ascertain whether their subsequent career preferences were influenced by the program or whether these preferences simply evolved from the student's previous interest in certain career paths. Self-selection is one of the most serious of all threats to the soundness of evaluation studies, and when the basis for self-selection is related to the nature of the program itself, as in the case of preceptorships, extreme caution must be used in interpreting results.

Another threat to a study's soundness is maturation, or the natural changes which occur over time. Evidence from one longitudinal study indicates that there are patterns of decision making on the part of students which develop as they proceed through their training. Studies which do not take account of these patterns and their timing are likely to confuse these naturally occurring patterns with program effects and to conclude that the latter are responsible for changes which would have happened anyway.

Selection-maturation interaction is a combination of the two problems just mentioned, and it involves the self-selection of individuals into a study who would change regardless of whether they had entered the program being evaluated. As Houston states, "This is a threat to the internal validity of inference whenever pre-experimental designs are used with long-term programs involving non-random selection of participants" (Houston, 1972, p. 60). In his usuage, internal validity refers to whether or not the program itself accounts for the outcomes observed. Pre-experimental designs are those which do not have sufficient controls to permit valid inferences regarding program effects.

The three design weaknesses just mentioned are perhaps the greatest threats to the internal validity of evaluation studies designed to assess the impact of preceptorship experiences on student career paths. Others may be present in specific instances, and these weaknesses will be discussed in the context of specific designs described in sections III and IV. In general, all threats to internal validity involve the possibility that the impacts or outcomes observed may be due to factors other than the program being evaluated. Control is the key to neutralizing these threats, whether they be threats involving self-selection, maturation, or both, or simply not taking account of specific variables for which evidence indicates their effects on career preferences. All of these threats are very salient in the evaluation of preceptorship program impacts, and all evaluation designs which do not attempt to introduce controls for these factors are seriously weak in their internal validity. In many cases this weakness is great enough that the studies or designs are of dubious utility in providing information about the impact of preceptorship programs on medical students' career preferences.

Another type of weakness to be considered here is external validity. This refers to the generalizability of the findings obtained from an evaluation study. This is separate from internal validity: a study may have solid internal validity while at the same time be lacking in external validity. In such a case, the results would only be useful for making statements about program effects in the specific situation and context existing in the original study. Threats to external validity include inadequate sampling of individuals and situational factors or variables. Both of these are important considerations. A sound, internally valid preceptorship evaluation conducted on one program at one school may not have a representative sample of students. It is obvious that different medical schools have different selection criteria for their students, and the students attending one school may be quite different from those at another school. The problem is that we cannot know whether a preceptorship program at one school (with a certain type of students) would have the same or different effects on (different) students at another school. A "successful" program at one school might be totally ineffective at another. This is a matter of serious concern if one is thinking of using the evaluation results from one program (or even two or three programs) as an indication of the effectiveness of preceptorship programs in general.

Situational factors are also a matter of concern for external validity. These include such things as program settings, length of the experience, time in a student's career development at which the experience occurs, and numerous other factors which are likely to vary across programs. The results of not controlling these factors is the same as that mentioned above: one cannot be sure that these idiosyncratic factors are not influential enough to result in different impacts among different programs.

For the policy maker, the problems posed by attempting to use the results of studies with weak designs (i.e., of questionable internal and/or external validity) are very serious. It is bad enough not to have the kind of information necessary for making policy decisions, but it is equally bad not to know whether the information one does have is good or bad, accurate or inaccurate, or true or false. In the face of uncertainty concerning the information which is available, it would at least be of some help to know what the probabilities are that the information is accurate.

Unfortunately, it is absolutely impossible, given our present state of knowledge, to make any statement about the probabilities that the findings returned by a weak design are accurate. The best we can say, if we know that the designs have not controlled for factors known to be of some influence in career preferences, is that the results are quite likely to be misleading. But we cannot say how misleading in a quantitative sense. Consequently, in seriously weak designs (i.e., those which have left many factors uncontrolled which are known to influence career paths), we may be in the situation of having to say that the information returned by such studies is no better than not having any information at all since the uncertainty about program effects in both cases is high enough to render the two situations equivalent for practical purposes. In other words, the mere availability of study results implies very little about the quality of that information; and this, unfortunately, is true even for published studies.

We now turn to a consideration of some of the designs which could be utilized to evaluate the effects of preceptorship experiences on career preferences. We begin by looking at some of the simpler designs, i.e., those which are initiated after the fact.

II. RETROSPECTIVE DESIGNS

Retrospective designs are those which attempt measurements and assessment of program effects after the population being studied has been exposed to the program. We shall first discuss some general weaknesses of this type of design. Then we shall look at some specific retrospective studies of impacts of preceptorships. Following that, we will consider some retrospective studies which could be undertaken using existing data. Finally, we examine some of the specific problems associated with the use of this type of design for evaluating the impacts of preceptorships.

General Weaknesses

In a general sense, any type of design is adequate and will yield accurate information about program effects so long as it enables one to rule out any plausible alternative explanations for the effects observed.

The problem, of course, is that some designs are much better at eliminating alternative explanations than others. As Campbell and Boruch put it, "There is, in particular, one plausible rival hypothesis that invalidates most social experiments of a quasi-experimental design nature: that is, there is a profound underlying confounding of selection and treatment" (Campbell and Boruch, 1975, p. 203). As was mentioned earlier, this is frequently the case in preceptorship programs, especially elective ones. Even with required programs, however, there is no way to distinguish between those who would have chosen to take the program and those who would not have taken it had it been an elective.

Houston is even more critical of retrospective designs which attempt post experience measurement only on the group exposed to a program:

> A pre-experimental design is a procedure for gathering observations which does not permit valid inference regarding program effects. While it may seem surprising that such procedures are used at all, they possess the advantage of requiring little effort, and are of some persuasive value among persons unfamiliar with scientific research. They often avoid the problem of statistical power entirely by making no provision for concluding that a program had no effect...
>
> The simplest pre-experimental design is the "post test only" design. A group is exposed to a program, and measures are subsequently obtained to characterize the group. No explicit comparison is made with any other group, this "case study" being presented as evidence of the program's effectiveness. While the evaluator may argue that the group is different from how it would have been without the treatment, our ignorance of how it was before the treatment gives us no basis for agreement or disagreement with his assertions....Such studies are woefully lacking in control, meaning, or scientific value (Houston, 1975, p. 59).

Not all evaluators and researchers are as critical of "pre", "quasi", or otherwise "non" experimental designs, but even those who are less critical recognize the consequences of using weaker designs. Cicirelli et al. point out that "To rule out ex post facto studies, for example, is to virtually guarantee that the use of social science research for policy decision-making will be even less than it is now." They go on to acknowledge some of the pressures that make it difficult to use "pure" designs and recognize the compromises that generally are made. They then add, however, that "...such compromises involve risks and the greater the compromises the more likely it is that results will be in error" (Ciricelli et al., 1970, cited in Rossi and Williams, 1972, pp. 182-183)

Retrospective studies which attempt to match "control" and treatment groups by identifying ex post facto those individuals who are like the treatment group but who were not exposed to the program are not as much of an improvement as they might seem. It is difficult enough to match individuals before a study (rarely is it possible to match them on more than three or more variables at most), but it is virtually impossible to match them after exposure to a program. The reason is that without a "before" measure of the treatment group, one cannot know what they were like prior to the program and therefore cannot know what a matched "control" group should be like. In fact, this shortcoming is so severe that it has been suggested that the term "comparison" should be substituted for "control" since there is in fact virtually no control at all in such a design.

Apart from design considerations, i.e., control of variables through specification of groups receiving or not receiving treatment, there is a serious problem of measurement in retrospective studies. By their very nature, these studies require that measurements be taken some time after the program or experience. This in itself is not necessarily bad if one is conducting a longitudinal study and if one assesses the necessary variables in an appropriate way. But if one requires individuals to recall what their reactions to treatment were at the time of the experience or to recall what they think the influence of the experience on them was, the probabilities of obtaining accurate measurements become quite small.

Evidence for the seriousness of this problem can be seen in studies cited by Deming in which individuals were being asked for factual information (e.g.,age). Second interviews 8 - 10 days after the initial questioning revealed that 10% of the ages from 8500 respondents were different by one year or more from the ages given initially. Deming suggests that other objective data with somewhat looser definitions (e.g.,education or employment status) would show even more variability (Deming, 1970, p. 324). If such problems are encountered with objective data on characteristics existing at the time of questioning, then the accuracy of data on subjective characteristics such as attitudes are preferences which the individual held a year or more previously is very questionable.

Retrospective Data Currently Available

A few schools have already conducted retrospective studies on their graduates or have data that could be used in a retrospective study. The Wisconsin study is fairly well known (Sivertson and Stone, 1972). It involved responses from 302 graduates from the classes of 1934, 1944, 1959, 1964, 1968, and 1970. Data collected included type of practice, location, and whether the preceptorship experience was influential in determining the graduate's career choices. The latter was obtained by asking each graduate whether he thought the experience influenced him. The conclusion made by the Wisconsin authors was that the study was "...most heartening for the proponents of the program. It has provided data on which to judge the merits of the program." They also stated that "...preceptorship does appear to influence career selection and does so in the direction of choosing a medical specialty. The fact that the preceptorship was an important influence on career strengthens the assumption that the close personal contacts with another physician do have a strong bearing on the student's choice of a career."

Missouri did a retrospective study of 1146 students (of whom 75% actually responded) who graduated between 1957 and 1973. It has not focused on ascertaining the effects of the preceptorship program, however. The data collected were for the most part "hard" data such as location of residency, hometown community size, community size in which practice is located, etc. While the results of this study, as yet unpublished, do not deal with preceptorship experiences, they do show that factors other than preceptorships are correlated with certain career choices. Such correlations do not prove causal relations among the factors studied, but they do indicate that we should be cautious about interpreting the effects of preceptorships in studies which make no provision for controlling other factors. This is expecially true for factors whose correlations with career preferences suggests that they may be as much responsible for the effects observed as the program itself.

Kansas also conducted a retrospective study of approximately 1000 graduates from the classes of 1951 through 1960. The response rate was 68%. Although the study was done to assess the preceptorship program, little in the way of career choice data was obtained. Type of practice (general, specialized, residency, military, etc.) and satisfaction with the preceptorship experience

were the primary variables of interest. No attempt was made to ascertain the effects of the experience on career pathways.

Despite the willingness of the Wisconsin study authors to draw conclusions from their data, neither that study nor the Missouri or Kansas studies provide the kind of data which warrant making firm inferences about effects of preceptorships. It is interesting to note that the authors of the Wisconsin study were both physicians. Physicians often work with retrospective data in the form of patient histories and apparently have few reservations about using similar data in social research. Their willingness to use such data has little relevance regarding its acceptability by standards of social research, however. The Wisconsin study exhibits all of the weaknesses of retrospective designs mentioned earlier. While it may suggest some hypotheses regarding preceptorship impacts on career dispositions, that study is not suitable for drawing any conclusions about these impacts. In short, all of these studies must be viewed primarily as raising questions rather than as providing a basis for making even weak inferences about the effects of preceptorships.

Potential Sources of Additional Retrospective Data

Data collected by other shcools could perhaps be used in retrospective analyses if money were available for such projects. The schools to be mentioned here have such data, but have not analyzed it in any systematic way. Possibilities for using these data in retrospective designs will be discussed, following which the most serious shortcomings of these potential studies will be described.

The Unversity of Illinois collects only narrative, impressionistic data from its preceptees and does not attempt to determine career preferences. Data are available on placements of graduates, however. Since the preceptorship program is an elective one, then assuming that academic records have been kept indicating students' choices of electives, one could compare the placements of preceptees with those who did not elect the experience. Such a design would be more than a simple post-experience retrospective study in that it would involve a "comparison" group. The University of Minnesota

has data similar to that of Illinois. Their program is elective, data are available on placements of graduates, and a similar retrospective design could be used there also.

Retrospective studies could also be done at a number of other schools besides Illinois and Minnesota. These schools include Michigan, Iowa, Ohio State, Utah, the Medical College of Virginia, and the NHSC program. These schools collect some kind of career preference data, and their students could presumably be studied retrospectively sicne these schools have all indicated that they "track" their students. Data available from these schools will be discussed elsewhere primarily because their data collection designs lend themselves to other types of evaluation studies. The earliest classes for which data are available from any of these latter schools is 1971.

Since both Illinois and Minnesota have elective preceptorship programs, one ostensible advantage of doing retrospective studies at these schools is that a comparison group would be available at both of them. These comparison groups would be those students who did not elect to take the preceptorship. Unfortunately, by their very nature elective programs confound self-selection and program effects, and this fact would invalidate such a study no matter how much the comparison groups were like the preceptees in other respects except exposure to the program. It would not be possible to make a valid inference that any differences in career preferences between the two groups were due to the program effects. In other words, comparison groups offer no advantage at all so long as selection and program effects are not separable in the study design.

It should also be pointed out that retrospective designs make it virtually impossible to assess and identify the specific features of programs which may be responsible for any program impacts. With some effort, data such as length of preceptorship, amount of financial support given to students, whether credit is given or not, etc. may be obtained about a program retrospectively, but these are probably not the most important features of programs as far as their impact on students is concerned. Other features of programs may contribute to the impact on career preferences, particularly those having to do with the psycho-social aspects of the experience. Without measuring

such characteristics in a detailed way across schools (so as to obtain enough variability to permit determination of the effects of these characteristics), one is placed in the position of being able to say (providing the data and design justify any conclusions at all) only that something called "preceptorships" had some kind of observed effects. Without knowing more about the characteristics of the programs it is not really possible to say just what it was that produced the observed effects, let alone to make any improvements so that programs might have greater impacts.

The issue just discussed is perhaps a moot point in retrospective studies since for the most part they do not justify drawing any sound conclusions, but it applies equally as well to more sophistocated kinds of studies. We now turn to some of these other types of designs.

III. CROSS-SECTIONAL DESIGNS

As used here, "cross-sectional" refers to those designs in which data are gathered close to the time at which the student has his preceptorship experience. These designs differ from retrospective studies, in which data are gathered a year or more after the experience, and from longitudinal designs, in which data are collected for extended periods before and after the experience. There are innumerable types of cross-sectional designs -- too many to be discussed even in general terms in this report. For this reason, selected cross-sectional designs will be discussed in some detail. Their descriptions should give some idea of the strengths and weaknesses which are encountered with this category of evaluation design.

General Weaknesses

There is one general weakness which is common to all cross-sectional studies which purport to determine program effects on career choices. This weakness derives from the fact that career development is a process occurring over a fairly extended period. Any cross-sectional design which focuses on the immediate experience of a preceptorship program cannot directly assess ultimate career decisions, which is, of course, the outcome of interest here in evaluating preceptorships. This problem has often been noted by other evaluators of social programs. Bennett and Lumsdaine state, for example, that....

> A further problem is the common tendency to assess programs in terms of things that are convenient to measure. An important special case is the difficulty of measuring this year the outcomes that are only going to eventuate several years hence (of which one has at best only a very unreliable prediction). We may have quite precise measures of immediate effects, but the longer range effects, which are less well measured if measured at all, may represent the outcomes that are the most important. This can place us in the strange position of having to weight indices of program impact in inverse proportion to the reliability with which they can be measured. (Bennett and Lumsdaine, 1975, p. 28).

Approaches to dealing with this probelm are not easy to come by. At best we may have outcome measures whose reliability we cannot estimate. At worst, we may find that the measures we used were wrong by as much as 50% or more. Not knowing what the outcomes of a program are is not a happy state to be in if one is trying to evaluate the program's impact, but this is the situation with cross-sectional studies of preceptorship effects on career choices. Cain and Hollister suggest one rather uncertain way of dealing with the problem.

> It is particularly difficult to predict long run or permanent behavioral changes from the short run indicators revealed by the ongoing or just completed program.
>
> The outcomes we wish to measure from many social action programs occur months or years after the participants have completed the program. We can use proxy measures, which can themselves be measured during and soon after the program, but follow-up studies are clearly preferred and may in many cases be essential. A good deal depends on the confidence we have in the power of our theories to link the proxies or short-run effects...with the longer run goals...(Cain and Hollister, 1972, p. 127).

Obviously when we measure a student's career preferences after he has been through a preceptorship, this is a "proxy" measure for his ultimate career choices, which may not occur for four or five years. Theory is not apt to be of much help since there has not been much theoretical work done in the area medical training and career choices, partly because federally funded research in this area has generally been limited to empirical studies. Still, it would be very helpful to be able to extrapolate from career preferences expressed after preceptorships to career choices made when students actually enter practice. If we know something about the attrition rates

from general practice preferences occurring between the junior year and entry into practice, then we might be able to estimate roughly how many students would actually enter primary care based on data obtained after preceptorships in a cross-sectional design. Unfortuantely, such attrition rates are not available on an inter-school basis to this author's knowledge.

The problem with cross-sectional studies, then, is that we cannot obtain measures of the career choices on which we think preceptorship experiences might have an effect. In those cases for which data are available on career choices (eg., Minnesota), we do not have measures of students' career preferences at the time of the preceptorship experience, and the study designs are not such as to eliminate alternative explanations for the choices actually observed.

Cross-Sectional Data Currently Available

Iowa conducted a cross-secitonal study of its classes of 1953 and 1954, and did a follow-up survey 10 years later. Assessments of career preferences were made before and after the preceptorship experience, including intent to enter general or specialized practice, preferred community size, and intent to practice in Iowa. Very little change was noted for specialty or general practice preference, and there was only an 8% increase in those indicating they would like to practice in a town of less than 10,000. There was an increase of 4% in those indicating they intended to practice in Iowa.

The follow-up survey 10 years later asked graduates whether the preceptorship experience had influenced them to enter general practice. The results were substantially negative; most said they felt the program had had no effect on their choice. At the time of the initial study, the ratio of those intending to enter general practice to those wanting to specialize was about 2:1. The follow-up survey, however, revealed that in terms of actual choices, the ratio had declined to 1:2, i.e., it had been completely reversed. To the author's knowledge, this is the only data available comparing undergraduate preferences for general practice with actual career choices. Data on community size location decisions was not reported, and geographic location was reported by state.

Ohio State has been evaluating their preceptorship program with a post-test only design and has data on some 800 students since 1971. It is not entirely clear whether all data are available on all students, and it is not known whether their students are tracked into practice. Their preceptorship is required, and thus no comparison group is available. Career preference data collected included the one most likely specialty choice, the one most preferred community size, and the preferred geographic region.

The three Michigan schools collect data yearly on students' career preferences, including the most likely specialty, most likely location by state and region, and most likely community size. It is not clear just when the annual data collection is carried out relative to the preceptorship experience, but it seems reasonable to presume that students are assessed at the same time. The preceptorship program is elective and lasts from 4 to 12 weeks. Given the assumption that students are assessed at the same time, one consequece is that the time period between student's preceptorship and data collection may vary from approximately 1 to 8 months.

Kansas assesses its students before and after the preceptorship. The primary focus on the evaluation is on conceptual learning in a variety of areas, although students are asked to indicate on a 5-point scale their degree of interest in entering primary care. This latter measure is the only career preference data obtained. The preceptorship is required, and data are available on the proportions of Kansas graduates entering various specialties.

Utah's evaluation design includes measures taken before and after the preceptorship. Career Preference data include rankings of the 5 most preferred specialties and work settings, and the one most preferred community size. Geographic area preferences are not obtained. The assessment instrument does include a check list of reasons for choosing a small or large community . The preceptorship at Utah is elective and about 43% of the students take it. The study was begun in 1972, but revision of the instruments since then may mean that not all data are available for all students. There is no continuous tracking of students, but a follow-up study is anticipated at an unspecified future date.

The Medical College of Virginia evaluates its preceptorship program with a post experience assessment. Data include degree of interest in Family Practice before the preceptorship (as recalled by the student), degree of interest in that specialty after the experience, one most preferred specialty, and community size preference. The remainder of the evaluation is concerned with various insights and understandings gained by the student. The program is elective and lasts 4 weeks.

Hershey evaluates their program with an assessment administered within 90 days following the experience. Career preference data are collected in the form of narrative descriptions by the student of his current specialty and location preferences, and a categorical rating of the degree to which he thinks the preceptorship experience affected these. The program is elective.

The evaluation design at the University of Illinois includes before, during, and after assessments, all of which are in narrative form. The program is elective, and data are presently available on placement of graduates. The narrative evaluation data is an unfortunate weakness in what would otherwise be one of the more thorough data collection efforts.

Wisconsin, the school with the longest continuously running preceptorship program, evaluates its program with a post-test only design. The career choice data obtained include a check list indicating whether the student thinks the experience influenced him to enter general practice, specialty practice, or neither; and whether or not it influenced his preference of a location. The latter item does not require the student to indicate either a specific community size or geographic location, but only whether or not the experience influenced his preferences with respect to these.

The NHSC preceptorship program administered by the AMAS Foundation assesses students before and after their elective experience. Data are obtained on a great many variables which have been found to be influential in determining career preferences, with the exception of institutional characteristics of the student's "home" school. The nature of the interaction between preceptor and preceptee is assessed as well as a variety of other experiential characteristisc. Career preference data include

ratings on a 7-point scale of preferences for a variety of specialties, community sizes, geographic regions, and work settings. Data on this program are still being collected and analyzed.

Evaluation of Existing Evaluations

Of the schools or programs for which data were available, those mentioned above are the only ones which presently appear to be collecting any kind of career preference data at all. Yet as Table 3 shows, not all of these programs collect even the minimum career data in which BHM is interested. Table 3 is quite instructive in the problems encountered in attempting to communicate what data are available. By a very liberal definition of the term, all of the programs listed in Table 3 can be said to collect career preference data. But as can be seen by the Table's footnotes, there are wide differences in the nature and scope of these data. Such diverse types of data, even when they can be lumped under the broad rubric of "career preference data" pose formidable problems for analysis and interpretation. These problems will be discussed briefly after a few specific design issues have been considered here.

In Table 4, the programs are grouped into different design categories. These categories represent the entire range of evaluation designs presently being used by programs for which information was available. Schools having elective programs are identified with an asterisk. If these shcools collect comparable data on their non-preceptees, then these programs have potential comparison groups which could be used in the evaluation design (except for the NHSC program). Utah and the Michigan schools are the only ones which are known to collect comparable data on non-preceptees; it is not certain whether the others do or not.

The five programs which collect only post experience data have the weakest of the cross-sectional designs. Generally speaking, it is quite risky to make any inferences concerning program effects without knowing what the students were like before the experience. Houston's comment about post-test only designs quoted earlier are an appropriate criticism. The weakness resulting from the absence of a pre-test is sometimes countered by having

a control group which is as much like the treatment group as possible (except that it does not receive the program experience.) However, without a pre-test, the only way to make the two groups alike is to assign individuals to the two groups randomly, and this is not done in any of the programs discussed here.

The four programs with before and after assessments represent stronger designs, but this strength is only relative. It is difficult to rule out alternative explanations beyond program effects which could account for any changes observed unless a control group is included in the design. Illinois has weak data (in narrative form) which would be very difficult to use in trying to detect measurable changes in career preferences. Of the before and after designs with a comparison group, this leaves only Utah. Thus, the Utah design appears to be the strongest, but it still has two significant shortcomings. First of all, data on preferred geographic area is not obtained, and second, the "control" group available is really not a control group at all. To quote Houston again,

> The experimental design which he (the evaluator) selects must permit the comparison of units exposed to the program with a control group which was not. The evaluator wishes to interpret measured differences between these groups as program impact; but a necessary condition for such inference is the absence of prior differences between the two groups. The burden of proof that no other differences existed lies heavy upon the evaluator. If, for example, his program were conducted in Boston, and his control group were found in San Diego, observed differences between the groups might also be attributable to the differential characteristics...of the two cites. Similarly, random individuals are not a good control for participants in a voluntary program, if the qualities which prompt one to volunteer could directly affect scores of the outcome measures. A group functions as an experimental control only if it eliminates the plausible alternatives to program impact in assigning a cause to observed outcome differences (Houston, 1972, p. 53).

Houston is not the only evaluation researcher with this view. Cain and Hollister state that

> Given the objective of the program, the question, "What difference did the program make?", should be taken literally; we want to know the difference between the behavior with the program and the behavior if there had been no program. To answer it, some form of control group is essential (Cain and Hollister, 1972, p. 114).

To put it concisely, the need for a control group is not an opinion. It is one of the principles of research in the social sciences. Differences of opinion may arise over the adequacy of a given control group, but even here the differences are likely to occur primarily between those who have training in research methods and those who have only a passing acquaintance with those methods. An adequate control group must be as much like the treatment group as possible when the study begins. So long as those entering preceptorship programs do so on an elective basis, any control group which one might devise is inadequate since all such groups would differ from the treatment group on the crucial variable of self-selection. Under such conditions, self-selection may be a more plausible explanation for observed effects than the program itself.

There are ways of eliminating alternative explanations, or at least making them less plausible. One of these will be discussed here in order to contrast a much stronger design wigh those presently in use. Only by this contrast can we obtain some perspective on the soundness of the information being returned by existing evaluation studies. As Gilbert et al. state,

> Randomized controlled field trials are currently our bast device for appraising new programs. The word "trial" suggests the direct comparison of the effects of treatments...In these studies, the experimental units -- individuals, families, school districts, whatever the unit -- are randomly assigned to treatments or regimens and carefully followed to find out what the effect of the regimen might be. The randomization helps in several ways. First, it avoids the dangers of self-selection or of biased selection...Second, it helps us control for variables we may not otherwise be able to control....
>
> "Controlled" refers to two matters: 1) that the choice of treatment for a site or an individual is primarily that of the investigator rather than the individual (once he agrees to participate)...2) that at least two treatments are being compared. "Randomized" refers to the use of chance at some stage to choose which units get which treatments (Gilbert et al., pp. 50-51).

The consequences of not using adequate control groups, i.e., non-randomized ones, are many, and two of the most serious will be mentioned here. The first is the uncertainty concerning the findings of the evaluation study. Gilbert et al., quoting from a report on the Emergency School Assistance Program, point out that...

> ....It is difficult to overstate the importance of the experimental design. Had there not been an experiment, we would have had to compare schools that were deemed worthy of ESAP funds to those that were not; no matter what statistical tricks we attempted to make the two groups comparable, the question of whether or not the differences we believed to be the result of ESAP were due to some other differences between the two schools would have always remained open... (Gilbert et al., 1975, p. 61).
>
> This uncertainty haunts nearly every attempt to reach firm conclusions from observational studies and nonrandomized field trials about the effects of social or medical programs, no matter how clever the analysts (Gilbert et al., 1975, p. 61).

After reviewing a number of medical studies on the effectiveness of various treatments and procedures, Gilbert et al. note a second consequence of inadequate evaluation studies which has rather strong practical implications. The point they make applies not only to medical studies but to social programs as well.

> Now that we have reviewed a number of nonrandomized studies, what overall conclusions emerge? One characteristic of nonrandomized studies that shows up in context after context is that even after repeated studies had consistently gotten the same result, there was always room to doubt the validity of the finding. Indeed, in several medical areas...the findings were not backed up with the same degree of investigator enthusiasm when put to the test of a carefully controlled randomized trial. This weakness of uncontrolled trials poses a particularly difficult problem for the policy maker faced with what appears to be more and more evidence in favor of a particular course of action, when he knows that he may just be seeing the same biases again and again.
>
> Nonrandomized trials may not only generate conclusions with doubtful validity -- they may actually delay the implementation of better evaluations. If a series of nonrandomized studies indicates that a certain treatment is highly effective, some investigators will point out the ethical difficulties of withholding this apparently valuable treatment from some persons in a randomized trial. Yet the evidence in this section suggests that nonrandomized studies may often artificially inflate our estimate of a treatment's value. Whenever this happens, and a randomized trial is postponed as a consequence, an ineffective treatment or program may be administered for years. The opportunity cost of such a mistake can be high (Gilbert et al., 1975, p. 134).

It is apparent that present evaluation designs in use at those schools which do collect career preference data fall far short of being able to provide us with results which we can accept with any degree of certainty.

Up to this point, our consideration of these designs has been restricted merely to their gross features, i.e., the nature of their comparison groups and testing periods. Yet there are other factors which must be taken into account also. Lacking randomized control and treatment groups, it would still be possible to obtain some idea of program impacts if we had measures of those variables which appear to affect career preferences. This latter approach would not yield results as strong as we would obtain with proper treatment and comparison groups, but it would give us more to go on than what we have seen up to this point among existing evaluations.

Unfortunately, most of the program evaluation efforts which were cited above do not collect data on the relevant variables for this approach. In their review of programs attempting to influence the spatial distribution of physicians, Eisenberg and Cantwell note that

> A thorough evaluation of physician location programs is presently hampered by a lack of necessary data. Analysis of specific programs would require, at minimum, data on the backgrounds of program participants, exact specifications of program operations and detail of eventual outcomes. Unfortunately, for most programs, whether at federal, state, or local levels, these types of data are not readily available. In many cases, they do not exist (Eisenberg and Cantwell, 1976, p. 456).

Although their focus of concern was not limited to preceptorship programs exclusively their comments are quite relevant. Based on the information available for this report, even the minimum data mentioned by Eisenberg and Cantwell is not collected in most preceptorship evaluation efforts. The strongest existing evaluation effort, that by Utah, does not include much in the way of program characteristic data. The NHSC effort assesses nearly all of the relevant variables at considerable length, and is the most complete in this respect, but its design lacks a comparison group.

To summarize, existing efforts at evaluating the effects of preceptorships on career paths have many major weaknesses. Among these are: confounding of self-selection and treatment or program effects; lack of adequate control groups; lack of measurement of other variables thought to influence career decisions; lack of appropriate outcome measures (only 3 programs measure the three outcome variables of interest, and not all of these measure them adequately); and finally, lack of direct measurement of the ultimate

outcomes of interest (preferences and intentions only are measured, not actual choices or decisions). It should be noted that these shortcomings are only the most blatantly obvious ones; finer points of evaluation methodology have not been considered here, partly because the major weaknesses cited are severe enough by themselves to raise strong doubts about the utility of the data returned by existing efforts.

It should be emphasized that the criticisms raised here are not directed at existing evaluation efforts per se. Rather the comments made here refer to the adequacy of these efforts for answering three rather specific questions; what are the effects of preceptorship experiences on medical students' specialty, geographic location, and community size choices? It should be recognized that most of the existing evaluation efforts were not designed with these questions in mind; most, in fact, appear to have been undertaken for a wide variety of other purposes. This should be obvious from the relatively small number of shcools which bother to collect career preference data at all (of the 35 programs from which instruments were available, only 11 collect anything that could be labeled "career preference" data).

That this is the case should really come as no surprise. As Cain and Hollister note, even when we know what variables are relevant to an evaluation effort, "...we often do not have measures of those we do know about. The latter problem frequently arises because of the unfortunate practice of inviting the evaluator in after the program has been run and data have been collected" (Cain and Hollister, 1972, p. 122). The present report is being made some 6 years after preceptorship programs had begun to receive federal money on a fairly large scale. Apparently little knowledge was available at that time to what would be needed in the way of an evaluation, or even what questions it would be appropriate to ask regarding the outcomes of such programs. The schools receiving funds for these types of programs apparently have been free to develop their own evaluation efforts to answer whatever questions they deemed were relevant for their own purposes. Given such circumstances, there is no reason at all to assume that the concerns of individual schools or programs would conincide with the interests of

federal funding agencies when the latter became interested in impacts on career choices some six years after large scale funding had begun.

Informational needs change, of course, and it is not always possible to know ahead of time what data will be needed six or even three years hence. Given that this is the case it is quite probable that the most effective and rewarding way to evaluate preceptorships would be the longitudinal approach, which could collect data capable of answering not only those questions posed above but many others (perhaps as yet unanticipated) as well. For roughly 8% of the federal funds allocated to medical preceptorships during the period 1970-76, a longitudinal tracking system could have been undertaken which today would be yielding data on the actual career choices (not just preferences expressed several years prior to entering practice) of both preceptees and non-preceptees and fairly sound data on the impacts of preceptorships on these choices.

Cross-Sectional Designs Involving Additional Data Collection: Some Considerations

We turn now to a consideration of other cross-sectional designs which could be used to evaluate the impact of preceptorship programs on career paths of medical students. No attempt will be made in this section to describe specific designs. Rather, some of the basic considerations which might be taken into account in the development of improved cross-sectional designs will be briefly sketched.

The first, and perhaps most basic consideration of all, is that there does not exist -- not by any stretch of the imagination or subtle twisting of definitions -- a "minimum" set of variables and design specifications which are necessary for providing answers to questions regarding impacts of preceptorships. No evaluation design can provide absolutely certain answers. Any reasonably competent methodologist could probably find weaknesses in any evaluation design and raise questions concerning the soundness of the study's results. Evaluation studies all vary with respect to the certainties of their results or the soundness of their conclusions. The question, then, should not be stated as "What is the minimum evaluation effort required for providing the answers sought?", but rather as "What is the level of certainty desired or needed for the evaluation results to be useful in decision-making?" As was mentioned in Section I, it is virtually impossible to put any numerical probabilities on the certainties of study results before the study is conducted. The best that can be done is to give some estimate, based on past experience, experience of other researchers, and knowledge of previous empirical results,

as to the relative seriousness of various types of deficiencies.

Some of these deficiencies will be discussed below and they are also shown in Table 5 where they are ranked in descending order or seriousness. It cannot be too strongly emphasized that the ranking given in Table 5 is not absolute, but is based on judgments which are derived from past experience in research in medical education and in physician career development. Not all methodologists would agree with the ranking, but most would agree that these deficiencies seriously threaten the soundness of any study's results. It should also be made clear that the seriousness of an evaluation design's deficiencies depends on factors beyond the five listed in Table 5. The unique context and peculiar circumstances in which a study is conducted may increase or decrease the impact of any single shortcoming on the soundness of the results. It is also true that particular combinations of the deficiencies listed in Table 5 may be more serious than one would expect from considering any one of them alone. These weaknesses will be discussed in the same order as they appear in Table 5.

1) The most serious deficiency has been indicated as self-selection. Persons who are interested enough to volunteer for a program or take it as an elective are likely to show changes in attitudes that reflect only the interaction of the experience with their prior interest. It is much easier to be "successful" at many things if one operates a program with persons who are already intersted in that program. This deficiency is serious not only for this reason, but also because "self-selection" as a variable is merely an indicator of a great many other variables which may affect an individual's score on any outcome measures. Self-selection may be an indicator of attitudes towards rural practice, primary care, or of degree of interest in practicing in a certain career niche, to mention only a very few of the variables which may become confounded with program effects. When this confounding occurs, it is virtually impossible to conclude that any outcomes observed were due to program impact.

2) Lack of an adequate comparison group also vitiates any conclusions which might be forthcoming from an evaluation study. Most studies seek to discover whether the program in questions made any difference, i.e., produced

effects which would not have occurred without the program. If one cannot compare the "treatment" group with a control group which did not recieve the treatment or program, there is very little basis for concluding that the program was effective. Such a comparison is essential for evaluating the results of a program. Without it, one is limited merely to saying (assuming that before and after measures were taken) that those receiving the treatment were different (or the same, as the case may be) after the program -- but it is not legitimate in this case to conclude that the program was responsible for the difference, if any.

The nature of the comparison group is vital also. It is not sufficient to have just any group serve as a comparison to the treatment group, especially when the program is an elective one. Ideally, the comparison and treatment groups should be as much alike as possible on crucial variables before the program is started. To the degree that these groups are different, there must be serious doubts that outcome differences resulted from the program's effects rather than from initial dissimilarities or interim developmental processes.

3) Without measures being taken before the program begins, it is absolutely impossible to conclude that the program had any effect if no comparison group is available which has been randomly chosen from the same pool as the treatment group. Again, the essence of the matter is to have a comparison of the treatment group outcome measures with something. Without some pre-experience measures, there is nothing to compare the outcome measures against, and hence no basis for making any conclusions regarding program effects.

4) There are two points to be made concerning the absence of measures of those "other" variables which inhere in the person, situation, or program and which may also influence the outcomes of interest. The first is that students' background characteristics have often been found to be predictors of their career preferences, and these characteristics must be taken into account in order to separate the effects of these variables from program effects. Without randomizating treatment and comparison groups it becomes necessary to measure these variables in some way. It is also important to

measure program characteristics for purposes of determining just what features of programs were responsible for observed effects, if any. Without this type of information, it is impossible to know how to maximize potential program impacts. Furthermore, given the diverse nature of the various existing preceptorship programs, this type of information is needed in order to know just what kind of program it was which produced the effect in the first place. This latter issue will be discussed further in the section dealing with the possibilities of compiling existing data.

The second point is that while regressions can be done in situations where only a treatment group exists, there are problems with this and other statistical techniques which may arise when a control group is lacking. Regression equations can only be based on the data actually observed. If data are gathered only from a treatment group exhibiting self-selection, then the variances in the predictor variables of the equation may be reduced from what they would have been if an adequate control group had been included. The result would be that estimates of the slope and weights in the regression equation may be far different from what they would have been with a sample which was not restricted to self-selectors.

5) The final weekness discussed here concerns the use of indirect measures of the ultimate outcomes of interest. Without knowing what specialty a student actually enters (or geographic area or community size), we must rely on his expressed preferences. In preceptorship evaluation studies these are often assessed before he even enters residency. If one simply wishes to know whether the student's career preferences have changed as a result of his preceptorship experience, such measures are perfectly adequate; but if one is interested in knowing whether that experience has affected the choices a student makes when he enters practice, then such measures will not provide the answers we are seeking. The seriousness of the discrepancy between expressed preferences of an undergraduate and the career choices he makes upon entering practice is not likely to be questioned by those familiar with research in medical career development. If the opinions of researchers are not sufficient, however, there is empirical evidence for corroboration. The Iowa 10-year follow-up study of their preceptees showed

that of those who indicated after their preceptorship that they intended to enter general practice, only half were actually in general practice 10 years later. The use of indirect indicators of ultimate outcomes is judged to be a less severe shortcoming than the first four listed in Table 5 since we at least have evidence of the direction in which these indicators are in error.

There are ways of compensating for the shortcomings listed in Table 5. Two points should be recognized, however. The first is that most, if not all, of the means of compensation can at best only partially correct the deficiency. Second, and perhaps even more significant, compensating for any given weakness listed in Table 5 may require eliminating one or more of the other weaknesses in the Table if they are present in the design. Lack of "before" measures, for example, is sometimes dealt with by having a sophisticated design of control and treatment groups. In short, compensating for a specific design weakness often means that we must improve other design characteristics as much as possible. It is apparent, however, that the existing designs discussed in this section do not have adequate compensating mechanisms. All of them confound self-selection and treatment, and all of them use only indirect measures of outcomes.

The point to be drawn from the discussion above is that any cross-sectional designs involving new data collection should eliminate as far as possible the weaknesses listed in Table 5. Any new design which does not attempt to compensate for these weaknesses would probably not be worth the effort since it would be no better than the evaluations presently being conducted.

IV. COMPILING DATA FROM EXISTING STUDIES

The focus in this section is on the possibilities of collecting data from current evaluation efforts and pooling it into one large data base which could be tapped to answer various questions regarding preceptorship program characteristics and impacts. Among the advantages accruing from such a data base would be: the cost would be lower than a new data collection effort;

data would be obtained across schools and programs; data would be obtained across graduating classes; the data would be collected in a central location where it would be available for answering various types of questions about preceptorship programs and limited questions about medical manpower; and the data could be handled by a central staff which would have responsibility for storage and accessing of the data, thus reducing by perhaps many months the time lag required when requesting information from several independent investigators. Unfortunately, while this arrangement might appear ideal (and in fact a similar system involving a new data collection effort was described in the original report under this contract), there are two major problems which would be encountered in attempting to create such a system from existing data.

The Problem of Comparability

The first problem is that of comparability and completeness of data. Evaluation studies which were implemented independently cannot be expected to have exactly comparable data. Yet for purposes of storing data from various programs in a computerized file, it is necessary to make the data comparable. To illustrate the nature of this task, Table 6 shows the kind of data collected by four programs on the three outcome variables of interest, viz., specialty, community size, and geographic region preferences. Several points are worth mentioning concerning the information in Table 6.

Making the data comparable across programs, where this is possible involves collapsing response categories so that these are the same. For the specialty variable, for example, Michigan's 54 specialties must be collapsed into Utah's 8, and the same must be done for Ohio State's categories. This collapsing is necessary because it is obviously not possible to expand a smaller number of response categories into a larger number since the information required to do so is simply not available in the smaller data set. Two consequences result. First of all, it means that a considerable amount of re-coding must be done before entering the data onto the common file. For open-ended response data, this coding must be done by hand. For specialties this should not be too difficult, but it would be time consuming. Secondly, and perhaps more seriously, the need to collapse data categories

means that each time this is done, information is lost. Furthermore, it means that for every variable it will be necessary to adopt the response categories being used by the program which collects the least information on that variable. Kansas, for example, only asks students to indicate how likely they are to enter primary care. Were data from that program to be included in the file, then information from all schools about student specialty preferences would have to be entered as "yes" or "no" regarding inclination to enter primary care. It might be possible to devise a sophisticated file structure which would permit one to retain some of the information which would otherwise be lost in collapsing response categories, but even so, when analyzing the data it would still be necessary to use those response categories containing the least information. While it might be possible to construct such a system, it would be no more advantageous than either utilizing data on a program-by-program basis or collapsing all the data as described above.

Another problem encountered in making the data comparable concerns the open-ended types of responses shown in the Hershey data. Not only would these data have to be hand coded for each student, but such coding would invariably involve implicit transformations of the data in ways that are unlikely to reflect the student's intended response. How does one categorize a community size response such as "Anyplace but Hitching Post, Kansas"? Not only would the hand coding of open-ended responses often distort the individual student's intent, but it would also require altering the more informative responses of other students. A student who responded to an open-ended community size item with "A non-rural community" would either require us to collapse all responses into rural vs. non-rural or to classify that student as "undecided."

It should be emphasized that the examples cited in Table 6 for the most part involve the most easily comparable data sets, except for Hershey, which was included for balance and contrast. Inclusion of data from additional programs would only make the difficulties worse, as can be seen by reviewing the footnotes to Table 3 or the questionnaires in the Appendix of this report. Table 3 also gives some idea of the amount of missing data which one could expect were an attempt made to create a common data base using existing data.

Whether or not the type of data comprises sketched above would provide an acceptable basis for decision-making must, of course, be up to the decision-maker. Katz's comments are appropriate here:

> ...its use (i.e., computer technology) in social systems may not prove faithful to its origins. The public pressure for accountability and the need of administrators for quantitative indicators come at a time when computing services are readily available. Hence, there is the danger that the complexities of adequate data gathering may be bypassed and any type of readily available or contrived data may be fed into the computer...
>
> The history of poor support for empirical research in the social sciences has aggravated the present problem. Social scientists, through lack of resources, in the past have been content with secondary analysis of data gathered by others for other than scientific purposes...In passing, I would venture the proposition that a field research project that does not devote a major part of its budget to data collection is in all probability poorly planned (Katz, 1975, pp. 467-68).

The point is not that raw data available from existing evaluation efforts are contrived, but rather that they were gathered by a diverse set of researchers for the purpose of answering a disparate set of questions. The contriving may occur when these different data sets are forced into some kind of unified system.

Identifying Program Differences

The second major problem in attempting to combine existing data is centered around the need to distinguish program differences when pooling the data from different programs. Edwards and Guttentag put the matter very well.

> Typically, an evaluation experiment will compare one program with another or with no program. But what is a program? In principle, anything you wish to label as a variable can be treated as one. But nonsense independent variables will, via any experiment, lead to nonsensical conclusions. And it certainly is nonsense to assume that some program, implemented in different ways by different people in different places, is a single entity simply because it is called by a single name and perhaps funded from a single source of money (Edwards and Guttentag, 1975, p. 445).

The difficulties of evaluating diverse programs which have been given the same name are immense. Williams states the issue in even more hopeless terms.

> (The Elementary and Secondary Education Act)...in deference to the states' prerogatives in the educational field simply made money available for helping the disadvantaged without any very strong stipulations as to how it would be used. The money dribbled into the educational system and mixed with other monies raises real questions about its effective use, and at the same time makes it almost impossible to measure that effectiveness in any kind of rigorous evaluative study (Williams, 1972, pp. 290-91).

He was dealing with a different case, but with preceptorships the situation is very similar. Medical schools have been given a wide leeway in what kinds of preceptorship programs they will implement, and in most cases have even set their own goals as they wished. It is without question and justified, then, that the preceptorship programs in operation exhibit diverse characteristics.

For purposes of trying to combine existing data into a common data base, the problem is that most programs do not collect adequate or similar data, if any, about the characteristics of their preceptorships. For administrative purposes, there is no need for them to do so -- program directors know what their programs are like, so why gather data about them? Since there has been no incentive for programs to attempt to compare their programs with others in any but informal impressionistic ways, there is no reason to think that such programmatic data would be available -- and it is not, at least not the kind that would be helpful in doing a national evaluation study.

There are ways around this problem, of course, and one of these would be to collect program characteristic data for those programs which were to be part of any pooled data base. The first task would be to determine which program characteristics were relevant. The NHSC instrument includes a fair number of these variables and would be a good beginning, but it requests information from students about their experience of the program. Use of this instrument as is would mean that one would have to begin the study with students newly entering the programs. Gross, administrative descriptors of programs are inadequate since they are not at the level at which the student actually encounters and experiences the program. Some compromise might be possible between these two extremes. At any rate, a new data collection effort would be necessary if programs are to be differentiated.

Without such differentiation, it is difficult to know just what one is evaluating.

From another prespective, however, it is not necessary to know how programs differ. With this other type of evaluation one in effect simply is asking "What are the effects of spending X dollars on something called preceptorships?" One is not asking, "Did this program make any difference?", or "Is this program more effective than that one?" or even "How can we maximize the effectiveness of the money allocated to preceptorship programs?". Rather, one is merely asking what the outcomes of certain line item expenditures in a budget are. This type of evaluation is merely mentioned here as an alternative approach which might be considered if other approaches turn out to be either unsatisfactory or impossible given existing fiscal and other constraints.

V. USE OF PACKAGED DESIGNS

The probelms discussed above in conjunction with attempting to pool existing data sources suggest another approach to evaluating preceptorship programs. Since existing data sources are so desparate as to make pooling quite difficult, one logical alternative is a system whereby various programs would yield comparable data which could then be compiled in a single data base. One way of doing this would be to prepare a preceptorship evaluation package which could be utilized by any number of programs.

The advantages of such an approach are obvious: comparable data; capability of comparing different programs; larger sample and ability to look more closely at certain specific relationships; and opportunity to overcome weaknesses of existing studies by creating a stronger design than any at present. The immediate disadvantages are also obvious: cost of designing such a package; cost of compiling data and hiring a central staff to analyze it; and the time delay before usable data would begin to come in. The cost of developing just such a package may be found estimated in the original report under Section IV-B, "Non-longitudinal Alternatives." The cost estimates which appear there do not include the estimates for distributing and promoting the package, however.

Apart from the disadvantages, there are several other considerations which should be mentioned regarding this approach. As can be seen from Table 3, there are only three existing evaluation studies which collect preference data on the three outcome variables of interest: Michigan, Ohio State and NHSC. Utah lacks only geographic area preferences and is otherwise a very well-designed protocol. It focuses primarily on clinical learning, and like the Michigan and Ohio State studies, collects little data on specific program or experiential characteristics. The NHSC study focuses heavily on experiential factors, but lacks clinical learning data. Thus, existing protocols considered separately may be somewhat inadequate, depending on the foci of interest. The Michigan instruments are shared among the three medical schools in that state; Utah's instruments are being considered for use by Arizona and Colorado; and the NHSC battery is being considered for use by University of Connecticut and other organizations similar to AMSA. These three sets of instruments, then, have gained acceptance by a total of 8 plus programs and would appear to be a good basis on which to build an evaluation package.

Depending on whether acceptance of an evaluation package is to be made voluntary, it would be well for such a package to include data on those areas which participating schools would find of interest. While the NHSC package was designed specifically with the interests of the Bureau in mind, many schools have broader concerns, particularly in the area of clinical learning. If usage of the evaluation package is to be voluntary, it might be advisable to assess this area as well. A new set of instruments combining some of the Utah items with those from the NHSC questionnaires would produce a stronger assessment package than any existing at present and would probably find wider appeal among medical schools than would either one of these packages singly.

There are problems with expecting voluntary adoption of evaluation packages even if these are made available without charge. The programs at each school are different, as are the research personnel, the objectives of the program, and the funds available for evaluation research. Under these circumstances it should not be expected that an evaluation package available without charge will be adopted without some changes being made by the persons using it at each school; and when this happens, the intent of distributing

a common package is subverted. Adding items to the package questionnaires would not be overly serious, but deletion or changing of items would. Experience with instruments developed in one longitudinal study indicate that instruments which are shared may undergo a metamorphosis when they reach another school.

It is also necessary that data collection designs at each school be as similar as possible. This means that if the package consists of pre- and post-measurements, then these measurements must be taken at the same times relative to students' preceptorship experiences. Preferably, before and after measurements should bracket the experience, with uniform time periods between testing and the experience across all schools. It would undermine the intent of the packaged evaluation effort to have some schools give students their post assessments immediately after the preceptorship, while others waited a month or more. At present, practices vary among schools in this respect, with some schools testing immediately after the experience, some permitting a time lag of as much as 90 days, and still others merely requiring that the questionnaires be turned before credit will be recorded for the experience. The reason that this matter is of concern is that comparability of data requires not only the same items on the questionnaire, but also comparable timing and conditions under which the assessment is made. Voluntary usage of packaged evaluation protocols would provide little incentive for such uniformity -- or even any means by which it could be achieved.

Voluntary use of a packaged evaluation protocol by programs also involves other hazards which may or may not occur, but which should be anticipated ahead of time. One of these is the fact that schools may use the packages differently, i.e., some may decide to measure students both before and after their preceptorships, while others may measure their students only after they have taken the program (perhaps for lack of manpower, funding for such undertakings, or even lack of interest or research sophistication). This latter situation would vitiate the utility of results from such schools and in fact would make it difficult to use the data in analyses involving the pooled before and after measures from other schools.

The most effective use of an evaluation package made available to schools would occur if schools having both required and optional preceptorships were included among the users of the package, as well as if "independent" preceptorship programs such as those administered by the AMSA Foundation and by the Rural Health Associates in Farmington, Maine were involved also. Including a variety of programs would make it possible to develop some evidence for the degree to which self-selection may confound existing results. Voluntary usage of the package would require that a special effort be made to obtain the cooperation of a range of programs varying in the manner just described.

A somewhat more difficult problem to deal with might occur in the compiling of the data in a centralized location for analysis. If participation is voluntary, there may be constraints placed by schools or programs on the use of their data by "outside" entities, and this may occur for a variety of reasons. One of the more important of these reasons is confidentiality. Programs simply may see the confidentiality of their students as a primary value and thus may not want to release their data in any form which would permit the identification of individuals. Such policies on the part of more than a few participating schools would mean that certain types of statistical analyses could not be done with data from these schools. Unfortunately these are apt to be the more powerful types of analyses requiring large numbers and capable of revealing patterns in the data which would be missed with simpler analyses. If it were not possible to use the data in ways which would permit making full use of the information contained in the data, then it would seem rather pointless to go to the effort of developing the package and distributing it. Consequently the problem of release and use of the data is one which should probably be dealt with through a preliminary survey of potential package users before the decision is made to develop such a package.

One way to detour around the probelm just mentioned is to have each program or school analyze its own data. This approach has problems, too, however. A large drawback is that no school would have large enough sample

sizes to perform the kind of multivariate analyses which would be possible with pooled data. Another problem is that without funding specific data analysis efforts at each school, it would be risky to assume that the data would be analyzed and results made available when they were needed by manpower agencies. This is particularly likely with programs which do not have the internal resources to fund a full-time research effort and where the researcher is involved in many other activities such as teaching. One does not have to be in research very long before becoming aware of the many laments among numerous researchers about their lack of time or money for analyzing all of their data. Some data bases may go untouched literally for years.

An alternative implied by the discussions above is that of making the use of an evaluation package mandatory, or at least making funding of preceptorship programs contingent upon the use of such packages. This would perhaps eliminate most of the drawbacks of the voluntary system, but its feasibility would depend on numerous factors outside of the research effort per se.

A compromise between the completely voluntary and coercive approaches was sketched in the original report, where a system was described whereby programs would use instruments designed and administered by a central staff and would share the usage of the data with that staff. That system was described in the context of a longitudinal study, but it could be used just as easily in a cross-sectional survey. If funding of programs was made contingent upon the implementation of some kind of evaluation effort (left unspecified) this might encourage schools to participate in such a system without the policy being too coercive.

To summarize the comments of this section briefly, it is thought that if an attempt is made to package an evaluation protocol, then that effort could either include a revising or combining of instruments from a very few existing efforts, or it could involve the development of an entirely new battery of instruments. It is thought that, with the exception of the NHSC effort existing instruments lack the kinds of items which would provide the

the data the Bureau is seeking. This should not be too surprising since the NHSC instrument is the only one which was devised specifically for the Bureau's needs under a contract. A combining of instruments would probably provide the kind of data which would be of interest both to programs and to federal agencies.

Despite the additional cost and time required to obtain data, the packaging of a protocol which could be used by several schools is seen as an alternative worth considering because of the greater strength it would have compared to existing evaluation studies.

VI. SOME GENERAL ISSUES IN SELECTING AN EVALUATION DESIGN

Table 7 presents a summary of existing evaluation designs and selected new designs which could be used for assessing the impact of preceptorships on medical student career paths. The table is not exhaustive; the longitudinal design described in the original report has not been discussed in this supplemental report and is not listed in Table 7. The weaknesses of each of the designs are sketched very briefly in the table. The seriousness of these various shortcomings has been more elaborately discussed in previous sections. The designs have been ranked in terms of overall strength. The same cautions mentioned in conjunction with Table 5, are also appropriate for Table 7, i.e., not all methodologists would necessarily agree with this ranking.

Since Table 7 summarizes most of the major options open for evaluating the impact of preceptorships on career paths, it is perhaps worthwhile at this point to discuss some of the issues which are relevant to the choice of an evaluation design. These comments are presented here for the purpose of shedding light on some of these issues from a researcher's perspective, which may or may not be similar to that of the policy-maker. It is recognized that no one perspective is apt to address all of the concerns which must be taken into account by the policy-maker who selects an approach to evaluation or decides whether or not to evaluate a program. But it is thought that some input from a researcher's perspective and experience might be helpful in making such decisions.

The dotted line in Table 7 represents, by the author's and the AMSA Foundation's judgment, a differentiation between those designs which are worth using as sources of information about the impact of preceptorship programs and those which return information of such dubious value as not to be worth undertaking. Another way of putting it is to say that the author or the AMSA Foundation would consider publishing studies such as those above the line, but not those below the line.

Obviously this line is an individual judgment. Other researchers would not necessarily agree, and the policy-maker might be willing to accept information of less certainty than the author or the AMSA Foundation would. There are bound to be differences over where the line should be placed since persons in different roles have different needs. The decision making needs of the policy-maker will be discussed shortly; for the moment we look at some ostensible objections to using good designs and reasons for accepting weak designs.

Reasons for Disagreements Involving Weak Designs

Disagreements arise, not so much over the soundness of the best designs, but over the meaning of the results of weak designs. As Gilbert et al. observe, "Though controversy about policy implications may of course persist, few controversies about the effects of new programs survive a series of carefully designed randomized controlled trials" (Gilbert et al., 1975, p. 176). In other words, with good methods the results may not be absolutely unequivocal, but they are based on sound enough evidence that there is a high degree of consesus as to the meaning of the results.

With weak designs, on the other hand, disagreements over interpretation of results occur more frequently, and the reason for this is simple: the results are simply not as certain. Where there is room for uncertainty, there is also room for disagreement. As was pointed out in previous sections, weak designs are those which do not eliminate alternative explanations for the results observed. With weak designs, then, it is possible to attribute a variety of interpretations to the results since a variety of alternative explanations have not been ruled out by the design itself.

Disagreements over the utility of some of the weaker designs may also occur from differences in backgrounds of researchers. This was apparent in the Wisconsin retrospective study, where physicians with only medical training conducted social research using methods acceptable in gathering patient histories. The literature on evaluation research also gives the impression that those whose careers oscillate between government agencies and academic positions tend to have more tolerant views of weak methods than do those researchers who have not had experience in the area of policy-making and policy decisions. This latter observation suggests that in policy-making there may be different standards for the quality of information which is utilized for decisions. We now turn to consider some of the reasons for these different standards and for objections to the use of sound evaluation designs.

One of the concerns which arises in evaluation efforts is whether or not impacts can be measured and whether or not designs adequate for measuring impacts are feasible. In summarizing discussions at the Batelle Seattle Research Center Symposium, Lumsdaine and Bennett noted that

> ...there seemed to be general agreement on the desirability of randomized experiments, or controlled field trials, as a means of unequivocally determining the dependence of a particular outcome on a given programmatic treatment. The concerns expressed revolve about 1) the sufficiency of this experimental approach, ...and 2) the feasibility of an experimental approach, both in terms of its technical and logistic requirements and in terms of the political and organizational environment in which it must be implemented...(Lumsdaine and Bennett, 1975, pp. 527-28).

Obviously students' career choices can be measured easily; the question, then, is whether the impact of preceptorships on these choices can be evaluated using sound methods. The advantages of sound methods are great, as Houston points out.

> 1. The (experimental model provides a specific inference regarding the existence of effects which can be causally attributed to program components and their interactions; estimates of the magnitude of these effects are also provided.

2. The stability of these inferences is known, being specified by the size and power of the statistical tests afforded by the model.
3. The generalizability of these inferences is known, being specified by the experimental design.
4. The internal validity of these inferences rests upon assumptions generally accepted by behavioral scientists. (This last point is not an appeal to authority, but indicates the cogency of evidence obtained using the model. There is no similar guarantee, for example, that anyone will give credit to pre-experimental designs.) (Houston, 1972, p. 62.)

Objections to Sound Designs

In spite of the recognized advantages of sound evaluation designs, there often remain objections arising from the situational contexts in which a study is to be carried out. Two of these objections revolve around matters of cost and time. It is often objected that "good" evaluation studies are too costly, and it is true that they cost more than the weaker methods. Whether or not these costs are excessive is another question, however. Gilbert et al. address these questions.

> Despite the advantages of randomly controlled trials, three objections often emerge: 1) randomly assigning treatments to sites or to people raises political problems...2) randomly controlled trials take too long; and 3) they are too expensive to conduct....
>
> How much should society be prepared to pay for an evaluation of a program that leads to firm and reliable inferences about how well the program is working?...The answer to a related question is clearer. How much should society be willing to pay for an evaluation that does not lead to firm inferences about how well a program is working? Not much. Yet for years we have been paying enormous amounts for evaluations of this second kind. Had we earmarked a small portion of the budget expended on these efforts for a few randomized field studies, we might have realized three benefits: first, we might have learned with greater reliability how well or how poorly a particular program was working; second, we might have spent less on overall evaluation expenses; and third, we might have been able to use the results of the field trials to improve the operation of the program, so that it could better serve its intended beneficiaries (Gilbert et al., 1975, pp. 159-60).

Williams, formerly on the staff of OEO's Office of Research, Plans, Programs, and Evaluation, comments that good evaluations require adequate funding so that results are not based on heroic (i.e., questionable) assumptions, excessive shortcuts are not required, and time is available for realistic planning. He also notes that

> The idea of a long time horizon for some evaluative activities,.. will be difficult for key officials to accept. Pressures on them to act quickly are tremendous. It would be a sterile exercise to push aside these political considerations. At the same time, if a realistic attitude toward the time required for evaluative research connot be developed, it is difficult to see how real progress can be made (Williams, 1972, p. 297).

The objection about time seems to be weakened given the lengthy periods that some programs have been in existence without any evaluation effort being undertaken.

Another objection to the better, controlled evaluation studies is that is is not practically or politically feasible to manipulate people in the ways required by such designs. Gilbert et al. address this issue also.

> When we object to controlled field trials with people, we need to consider the alternatives -- what society actually does. Instead of field trials, programs currently are instituted that vary substantially from one place to another, often for no very good reason. We do one thing here, another there in an unplanned way. The result is that we spend our money, often put people at risk, and learn little. This haphazard approach is not "experimenting" with people; instead, it is fooling around with people. It is most important that we appreciate the difference between these two ideas. We need to decide whether fooling around with people's lives in a catch-as-catch-can way that helps society little is a good way to treat our citizens, and we need to understand that in current practice this is the principal alternative to randomized controlled field trials (Gilbert et al., 1975, p. 150).

Although their statement is rather strongly worded, it does raise an interesting question as to why it is apparently possible to manipulate people by developing and implementing programs but not possible to assign them to programs in ways which would give us sound information about what we are doing to them.

The quotations above have been concerned with randomized field studies, which are seen as the ideal and the only means of being strongly certain about program impacts. The present report has not considered randomized studies, primarily in recognition of some of the political difficulties involved in changing the ways in which students are assigned to programs or the way in which schools are funded after federal support for preceptorship programs has been in effect for several years. The studies above the dotted

line in Table 7 and the longitudinal design described in the earlier report are among the best designs that could be implemented without randomization. The weaknesses of these studies are such that their results would be good only for the populations on which they focus, viz., students who elect to take preceptorships. That is, we would not really have any basis for making statements about the effects of preceptorships on students in general, although the data from these designs would give us some clues in this respect.

Improper Use of Weak Designs

The decision as to which design returns acceptable data on which to base policy or funding decisions rests with those who must use results of evaluation studies. The role of the methodologist must be limited to reporting on the strengths of various designs, and it may be that the results of one of the weaker designs are adequate for the needs of the policy-maker. It should be clear by now, however, that the designs in Table 7 are not equivalent in what they do and cannot be considered on a cost basis as representing high and low "bids" for the same product. The designs listed in that table are very different entities yielding very different informational returns. One final caution is in order if any of the weaker designs are to be used.

Evaluation results are only one of the considerations that go into policy decisions regarding programs. As Williams observes,

> ...some caution is needed in interpreting evaluation data which generally will mean fitting the evaluation evidence into a mosaic with other reasonable evidence to make a decision. In general, the present generation of outcome evaluations should be viewed as a piece of evidence, not the definitive piece of information that bowls over all other reasonable indications of a different policy decision (Williams, 1972, p. 297).

The problem with weak evaluation designs is that their evidence is uncertain enough that innumerable questions can be raised as to the meaning of the results. Given a weak design, in all probability such questions will be raised when the results come in. What happens then is that the gray areas of uncertainty which cloud the results permit a variety of other concerns to enter the decision process under the guise of "scientific evidence."

These political considerations are a part of the decision process, as Williams points out, but they should be considered in and of themselves, not as they may enter the fray under the disguise of research results.

To employ weak evaluation designs is to invite just such an eventuality, and the result is equally as bad for those who attempt to employ the uncertain findings for their own purposes as it is for social research. Allowing political debates to creep into those gray areas of uncertainty created by weak study designs weakens the credibility of social research. The consequences are that those who misuse research findings in this way quickly wear out a tool which should not have been in the political arena in the first place, and the sound, useful results which could have been forthcoming from the social research community via stronger, sounder designs are lost due to an increasing decay in credibility. In this respect social research is much like any other tool in the sense that misuse of it may render it useless for those tasks and purposes for which it is appropriate.

Campbell and Boruch address this issue, and they point out that until organiziations have solved the problems of how to utilize the results of evaluation studies in non-political ways while at the same time giving them their proper weight among all of the other factors that enter into policy decisions, then

> ...evaluation methodologists should often advise <u>against</u> program evaluation, and should certainly not force it on unwilling institutions. Since true experiments are less equivocal than quasi-experiments, and since they permit fewer excuses for unwanted results, this legitimate fear of evaluation may often lead true experiments to be more opposed than quasi-experiments. Probably the correct stance will often be to recommend no evaluation at all... (Campbell and Boruch, 1975, p. 208).

Their comments and those made earlier are the reasons for the placement of the dotted line in Table 7 indicating which types of studies are viewed by the author and AMSA Foundation as not worth doing.

VII. OTHER TYPES OF IMPACTS WHICH COULD BE CONSIDERED

As stated at the beginning of this report, the main considerations regarding program impacts were limited to specialty, geographic area, and

community size preferences of medical students. This section will briefly discuss some other impacts of preceptorship programs which could be studied. It will also examine one more general issue regarding evaluation which is applicable to a variety of possible impacts.

Depending on the nature of the experience for the student and of the program itself, preceptorships may have additional impacts on students' careers beyond specialty, geographic area and community size preferences. Experiences may likewise affect student preferences regarding practice models, working with other allied health personnel, working with community boards, or even working with certain types of patients. Besides affecting their career plans, preceptorships may provide significant clinical learning experiences which students would not otherwise obtain. Although a tabulation was not made, it appeared from data available for this report that many more schools collect clinical learning data than collect information about career preference changes.

Other foci for evaluating preceptorships might be their impacts on the quality and/or quantity of health care delivered in the areas served, their cost effectiveness in providing mechanisms for continuing education, their effectiveness as a communications network for disseminating current medical knowledge and findings, and their impacts on other parts of the medical school curriculum.

Obviously not all of these types of impacts can be assessed using the designs or variables described in the presnt report. It may also be that the impacts in any one of the areas mentioned is relatively small. Any one kind of impact alone might not be of sufficient magnitude to justify continued funding of preceptorships, but taken together, the sum of all the various impacts might be more than enough justification for this type of program. This last observation brings us to one final point concerning the use of strong versus weak evaluation designs.

Because of the ambiguity of weak designs, small effects may not even be detectable with such quasi-experimental studies, and yet, as Cain and Hollister observe,

> Indeed, it is precisely programs which have large and dramatic effects which can be evaluated with a loose design and an almost journalistic level of evaluation, but we would contend that almost all social action programs, and particularly those in the field of manpower training and education, are unlikely to bring about spectacular changes (Cain and Hollister, 1972, p. 121).

Gilbert et al. continue the discussion by noting that

> The observation that many programs do not have slam-bang effects stresses the importance of measuring small effects reliably...Once small effects are found and documented, it may be possible to build improvements upon them. The banking and insurance businesses have built their fortunes on small effects -- effects the size of interest rates... Naturally, small effects require stronger methods for their accurate detection than do large ones. One must be sure that the observed effects are not due to initial differences between groups, or to other spurious causes (Gilbert et al., 1975, p. 177).

They go on to say that when small effects are observed, well-designed studies are essential for eliminating such sources of bias as alternative explanations.

If one really wants to know what some of the impacts of preceptorships are, it is apparent that sound designs are necessary for determining both the magnitude and sources of these impacts.

References

Bennett, C.A., and Lumsdaine, A.A., "Social Program Evaluation: Definitions and Issues," in Carl A. Bennett and Arthur A. Lumsdaine (eds.) Evaluation and Experiment, New York: Academic Press, 1975.

Cain, G.C., and Hollister, R.G., "The Methodology of Evaluating Social Action Programs," in Peter H. Rossi and Walter Williams (eds.), Evaluating Social Programs, New York: Seminar Press, 1972.

Campbell, D.L., and Boruch, R.F., "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives," in Carl A. Bennett and Arthur A. Lumsdaine, Evaluation and Experiment, New York: Academic Press, 1975.

Cicirelli, V.G., et al., "The Impact of Head Start: A Reply to the Report Analysis," Harvard Educational Review, 40, 1970, 105-109, cited in McDill, E.L., McDill, M.S., and Sprehe, J.T., "Evaluation in Practice: Compensatory Education," in Peter H. Rossi and Walter Williams (eds.) Evaluating Social Programs, New York: Seminar Press, 1972.

Deming, W.E., "On Errors in Surveys," in Norman K. Denzin (ed.), Sociological Methods, Chicago: Aldine, 1970.

Edwards, W., and Guttentag, M., "Experiments and Evaluations: A Reexamination," in Carl A. Bennett and Arthur A. Lumsdaine, Evaluation and Experiment, New York: Academic Press, 1975.

Eisenberg, B.S., and Cantwell, J.R., "Policies to Influence the Spatial Distribution of Physiciains: A Conceptual Review of Selected Programs and Empirical Evidence," Medical Care, 14, 1976, 455-468.

Gilbert, J.P., Light, R.J., and Mosteller, F., "Assessing Social Innovations: An Empirical Base for Policy," in Carl A. Bennett and Arthur a Lumsdaine (eds.) Evaluation and Experiment, New York: Academic Press, 1975.

Hilton, E.A., and Lumsdaine, A.A., "Field Trial Designs in Gauging the Impact of Fertility Planning Programs," in Carl A. Bennett and Arthur A Lumsdaine, Evaluation and Experiment, New York: Academic Press, 1975.

Houston, T.R., Jr., "The Behavioral Sciences Impact-Effectiveness Model," in Peter H. Rossi and Walter Williams (eds.) Evaluating Social Programs, New York: Seminar Press, 1972.

Katz, D., "Feedback in Social Systems: Operational and Systemic Research on Production Maintenance, Control and Adaptive Functions," in Carl A. Bennett and Arthur A. Lumsdaine (eds.) Evaluation and Experiment, New York: Academic Press, 1975.

Lumsdaine, A., and Bennett, C.A., "Assessing Alternative Conceptions of Evaluation," in Carl A. Bennett and Arthur A. Lumsdaine (eds.) Evaluation and Experiment, New York: Academic Press, 1975.

Otis, G.D., Quenk, N., Weiss, J., Albert, M., Offir, J., and Richardson, D., Medical Specialty Selection: A Review and Bibliography, DHEW Publication No. (HRA) 75-8, U.S. Department of Health, Education and Welfare, Public Health Service, Health Resources Administration, 1974.

Sivertson, S.E., and Stone, H.L., "Is the Preceptorship an Anachronism?" JAMA, 221, 1972, 590-592.

Weiss, C.H., "Policy and Strategy in Evaluation Research," in Elmer L. Streuning and Marcia Guttentag (eds.) Handbook of Evaluation Research, Vol. I, Beverly Hills: Sage, 1975.

Williams, W., "The Capacity of Social Science Organizations to Perform Large-Scale Evaluative Research," in Peter H. Rossi and Walter Williams (eds.) Evaluating Social Programs, New York: Seminar Press, 1972.

TABLE 1

List of Schools and Programs From Whom Information on Preceptorships Was Received

	School or Program Name	Did Not Respond	Program or Evaluation Not In Progress	Students are Tracked	Instruments are Available	Data on Career Preference
1.	Comprenehsive Health Service		x			
2.	Albany					
3.	Arizona					
4.	Baylor					
5.	Bowman Gray				x	
6.	Colorado			x		
7.	U.C. Davis				x	
8.	U.C. Irvine					
9.	UCLA		x			
10.	U.Connecticut					
11.	Cornell		x			
12.	Dartmouth		x			
13.	Duke		x	x		
14.	U. Florida				x	
15.	Frontier Nursing					
16.	Hahnemann		x			
17.	Hawaii				x	
18.	Hershey			x	x	x
19.	U. Illinois			x	x	x
20.	Iowa			x	x	
21.	Iowa (Osteopath)				x	
22.	Johns Hopkins		x			
23.	Indiana					

Note: Tracking of students can mean anything from yearly data collection to the fact that the alumni office maintains records of student addresses.

	School or Program Name	Did Not Respond	Program or Evaluation Not in Progress	Students are Tracked	Instruments are Available	Data on Career Preferences
24.	Kansas			x	x	x
25.	K.C. Osteopath			x		
26.	Kentucky					
27.	Kirksville Osteo.					
28.	Farmington, ME		x			
29.	L.S.U.					
30.	Maryland				x	
31.	Massachusetts				x	
32.	Mayo					
33.	Michigan*			x	x	x
34.	Mich. State Osteo.				x	
35.	Minnesota			x	x	
36.	Minn. AHES					
37.	U. Missouri			x	x	
38.	UMKC					
39.	N.J. Coll. of Med.					
40.	Mt. Sinai				x	
41.	U. Nevada, Reno					
42.	NHSC (AMSA Foundation)			x	x	x
43.	North Dakota				x	
44.	Med. Coll. of OH				x	
45.	Ohio State			x	x	x
46.	Oklahoma			x		
47.	Oklahoma Osteo.					

*Includes Wayne State, Michigan State and Univ. of Michigan

	School or Program Name	Did Not Respond	Program or Evaluation Not in Progress	Students are Tracked	Instruments are Available	Data on Career Preferences
48.	U. PA			x	x	
49.	U. Pittsburgh				x	
50.	Rochester					
51.	U.S.C.					
52.	So. IL		x	x		
53.	South Dakota		x			
54.	Stanford				x	
55.	SUNY Stonybrook				x	
56.	Temple		x			
57.	U.TX Galveston					
58.	U. TX Dallas				x	
59.	U., TX San Antonio		x	x		
60.	Utah			x	x	x
61.	Vermont			x	x	
62.	U. Virginia					
63.	VA Med. Coll				x	x
64.	WAMI**			x	x	x
65.	Wisconsin			x	x	x
66.	SNMA					
67.	IHS (Portland)					
68.	New Mexico			x		
69.	Albert Einstein	x				

**Program is not really a preceptorship

	School or Program Name	Did Not Respond	Program or Evaluation Not in Progress	Students are Tracked	Instruments are Available	Data on Career Preferences
70.	SUNY Syracuse	x				
71.	U. North Carolina	x				
72.	U. Cincinnati	x				
73.	Brown	x				
74.	Meharry	x				
75.	West Virginia	x				
76.	Med. Coll. of WI	x				
77.	Yale	x				
78.	U. Minn. Duluth	x				
79.	South Carolina	x				

TABLE 2

Tabulation by Various Categories of Schools Listed in Table 1

	Number
1. Total number of inquiries	81
a. No response	11
b. No evaluation or program not yet under way	12
c. Program is not a preceptorship	1
d. Subtotal, schools with no usable data	24
2. Total number of inquiries with usable, relevant data	57
3. Total number of schools who track their students	23
4. Total number of schools/programs for which instruments were available	35*
a. Schools/programs tracking students from whom instruments were obtained	17
5. Total number of schools/programs whose instruments revealed the collection of career choice data	11

*This figure includes those schools which indicated that they were sharing or borrowing instruments from another school.

TABLE 3

Descriptions of Data Obtained by Those Evaluation Studies Which Collect Career Preference Information

Programs	Specialty	Geographic Area	Community Size
1. Hershey	x	x	
2. U. Illinois			
3. Michigan	x	x	x
4. NHSC (AMSA Foundation)	x	x	x
5. Ohio State	x	x	x
6. Utah	x		x
7. Med. Coll. of Virginia	x		x
8. Wisconsin	x	x	
9. Kansas	x		

Note: x = these variables are collected in a quantitative, codable form.

Comments: (Numbers below correspond to numbers of programs in table above.)

1. Hershey asks for students' narrative remarks about specialty and geographic area preferences; students are then asked to give a quantitative rating of the degree to which their preceptorship experience influenced these preferences.

2. Illinois asks students for a narrative description of any changes they have made in their career goals as a result of their preceptorship experience.

3. Michigan asks students to indicate their most preferred specialty, area and community size and to indicate the degree of certainty of these preferences on a 4-point scale. Students without a strongest preference are asked to rank their first three preferences of a specialty.

4. Students indicate their preferences on a 7-point scale regarding each of 26 specialties, 6 geographic regions of the country, and 5 different community sizes.

5. Students report their first preference on each of these three variables.

6. Students indicate their most preferred among 6 specialty areas of medical practice, or rank their preferences if more than one is being considered, and pick one of 6 given community sizes in which they would like to practice.

7. Students specify the one specialty they are most likely to enter and rate on a 5-point scale their degree of interest in FP before and after their preceptorship experience, and select their most preferred of 6 community sizes.

8. Students indicate whether the experience influenced their inclination to enter general or specialized practice and their location preferences. Specific locations and specialties are not asked for.

9. Kansas asks their students to indicate on a 5-point scale how strong their preference is for primary care and to discuss narratively their reasons for these feelings.

TABLE 4

Evaluation Designs of Programs Collecting Career Preference Data

Pre- and Post-Test	Post-Test Only
NHSC	Ohio State
Kansas	Medical College of Virginia*
Utah*	Hershey*
Illinois*	Wisconsin
	Michigan* (3 schools)

*Indicates programs which are elective and thus potentially have a comparison group available. Only Utah and Michigan are known for certain to collect comparable data on non-preceptees, however.

TABLE 5

Selected Deficiencies in Evaluation Design Rank Ordered by Degree of Seriousness

1. Self-Selection
3. Lack of an Adequate Comparison Group
3. Lack of Pre-experience Measures
4. Absence of Measures of Other Variables Known to Influence Career Preferences
5. Proxy Measures of Outcomes (Undergraduate preferences rather than actual career choices)

TABLE 6

Type of Data Collected by Four Different Programs on Three Outcome Variables

Variables	Michigan	Utah	Ohio State	Hershey
Specialty	Check-list of 54 different specialties; student checks most preferred	Student indicates most preferred of 8 specialty categories	Student indicates most preferred of 16 different specialty categories	Open-ended response to most preferred specialty
Community Size	Student checks most preferred from list of: Metropolitan inner city " city " suburb City/non-metropolitan Town/non-metropolitan Rural community Rural remote Other	Student checks most preferred of: over 100,000 50,000 - 100,000 25,000 - 50,000 10,000 - 25,000 5,000 - 10,000 less than 5,000	Student checks most preferred of: rural area (less than 2,500) small town (2,500 - 9,999) large town (10,000 - 49,999) small city (50,000 - 99,999) large city (100,000+)	Open-ended, i.e., student can make own categories
Geographic Area	Student checks preferred state; states are categorized by region	Variable is not measured	Student is asked to write in state if it is other than Ohio	Variable is not measured.

TABLE 7

Major Design Options for Evaluating Preceptorship Impacts
(Ranked in Descending Order of Soundness)

Design	Existing Examples	Weaknesses
1. New Evaluation Package described in text	Described in detail in original report	1,2,5
2. Cross-sectional; comparison group before and after measures; follow-up	Iowa	1,2,3
3. Cross-sectional; comparison group; before and after measures	Utah, Illinois**	1,2,3,5
4. Cross-sectional; comparison group; post test only	Michigan, Hershey** Medical College of VA**	1,2,3,5,6
5. Cross-sectional; no comparison group; before and after measures	Kansas, NHSC	1,2,3,4*,5
6. Cross-sectional; no comparison group; post test only	Wisconsin, Ohio State	1,2,3,5,6
7. Pooling Existing Data	All programs listed in Table 3	1,2,3,4,5,6,8
8. Retrospective; one group; post test only	Wisconsin, Missouri Kansas	1,2,3,6,7

*This weakness is present for Kansas only.

**These schools may or may not have data on their comparison groups.

Note: All designs, including those with comparison groups, have been indicated as having inadequate control groups. This is because none of the comparison groups really meet the definition of a control group.

Key for Weaknesses:

1 = Confounding of self-selection and program effects

2 = Lack of adequate control groups

3 = Lack of measurement of other variables throught to influence career decisions

4 = Lack of all three outcome measures of interest

5 = "Proxy" measures of the ultimate outcomes of interest

6 = Lack of "before" measures

7 = Data of questionable validity

8 = Loss of information contained in original data